Chapter 3

Local Regression

Local regression is used to model a relation between a predictor variable and response variable. To keep things simple we will consider the fixed design model. We assume a model of the form

$$Y_i = f(x_i) + \varepsilon_i$$

where $f(x)$ is an unknown function and ε_i is an error term, representing random errors in the observations or variability from sources not included in the x_i.

We assume the errors ε_i are IID with mean 0 and finite variance var$(\varepsilon_i) = \sigma^2$.

We make no global assumptions about the function f but assume that locally it can be well approximated with a member of a simple class of parametric function, e.g. a constant or straight line. Taylor's theorem says that any continuous function can be approximated with polynomial.

3.1 Taylor's theorem

We are going to show three forms of Taylor's theorem.

- This is the original. Suppose f is a real function on $[a, b]$, $f^{(K-1)}$ is continuous on $[a, b]$, $f^{(K)}(t)$ is bounded for $t \in (a, b)$ then for any distinct points $x_0 < x_1$ in $[a, b]$ there exist a point x between $x_0 < x < x_1$ such that

$$f(x_1) = f(x_0) + \sum_{k=1}^{K-1} \frac{f^{(k)}(x_0)}{k!}(x_1 - x_0)^k + \frac{f^{(K)}(x)}{K!}(x_1 - x_0)^K.$$

  Notice: if we view $\sum_{k=1}^{K-1} \frac{f^{(k)}(x_0)}{k!}(x_1 - x_0)^k$ as function of x_1, it's a polynomial in the family of polynomials

$$\mathcal{P}_{K+1} = \{f(x) = a_0 + a_1 x + \ldots + a_K x^K, (a_0, \ldots, a_K)' \in \mathbb{R}^{K+1}\}.$$

- Statistician sometimes use what is called Young's form of Taylor's Theorem:

  Let f be such that $f^{(K)}(x_0)$ is bounded for x_0 then

$$f(x) = f(x_0) + \sum_{k=1}^{K} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k + o(|x - x_0|^K), \text{ as } |x - x_0| \to 0.$$

  Notice: again the first term of the right hand side is in $\mathcal{P}_{K+1}$.

- In some of the asymptotic theory presented in this class we are going to use another refinement of Taylor's theorem called Jackson's Inequality:

  Suppose f is a real function on $[a, b]$ with K is continuous derivatives then

$$\min_{g \in \mathcal{P}_k} \sup_{x \in [a,b]} |g(x) - f(x)| \leq C \left(\frac{b-a}{2k}\right)^K$$

  with $\mathcal{P}_k$ the linear space of polynomials of degree k.

3.2 Fitting local polynomials

We will now define the recipe to obtain a loess smooth for a target covariate x_0.

The first step in loess is to define a weight function (similar to the kernel K we defined for kernel smoothers). For computational and theoretical purposes we will define this weight function so that only values within a *smoothing window* $[x_0 + h(x_0), x_0 - h(x_0)]$ will be considered in the estimate of $f(x_0)$.

Notice: In local regression $h(x_0)$ is called the span or bandwidth. It is like the kernel smoother scale parameter h. As will be seen a bit later, in local regression, the span may depend on the target covariate x_0.

This is easily achieved by considering weight functions that are 0 outside of $[-1, 1]$. For example Tukey's tri-weight function

$$W(u) = \begin{cases} (1 - |u|^3)^3 & |u| \leq 1 \\ 0 & |u| > 1. \end{cases}$$

The weight sequence is then easily defined by

$$w_i(x_0) = W\left(\frac{x_i - x_0}{h(x)}\right)$$

We define a window by a procedure similar to the k nearest points. We want to include $\alpha \times 100\%$ of the data.

Within the smoothing window, $f(x)$ is approximated by a polynomial. For example, a quadratic approximation

$$f(x) \approx \beta_0 + \beta_1(x - x_0) + \frac{1}{2}\beta_2(x - x_0)^2 \text{ for } x \in [x_0 - h(x_0), x_0 + h(x_0)].$$

For continuous function, Taylor's theorem tells us something about how good an approximation this is.

To obtain the local regression estimate $\hat{f}(x_0)$ we simply find the $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$ that minimizes

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^3} \sum_{i=1}^{n} w_i(x_0)[Y_i - \{\beta_0 + \beta_1(x - x_0) + \frac{1}{2}\beta_2(x - x_0)\}]^2$$

and define $\hat{f}(x_0) = \hat{\beta}_0$.

Notice that the Kernel smoother is a special case of local regression. Proving this is a Homework problem.

3.3 Defining the span

In practice, it is quite common to have the x_i irregularly spaced. If we have a fixed span h then one may have local estimates based on many points and others is very few. For this reason we may want to consider a nearest neighbor strategy to define a span for each target covariate x_0.

Define $\Delta_i(x_0) = |x_0 - x_i|$, let $\Delta_{(i)}(x_0)$ be the ordered values of such distances. One of the arguments in the local regression function `loess()` (available in the modreg library) is the `span`. A span of α means that for each local fit we want to use $\alpha \times 100\%$ of the data.

Let q be equal to αn truncated to an integer. Then we define the span $h(x_0) = \Delta_{(q)}(x_0)$. As α increases the estimate becomes smoother.

In Figures 3.1 – 3.3 we see loess smooths for the CD4 cell count data using spans of 0.05, 0.25, 0.75, and 0.95. The smooth presented in the Figures are fitting a constant, line, and parabola respectively.

Figure 3.1: CD4 cell count since seroconversion for HIV infected men.

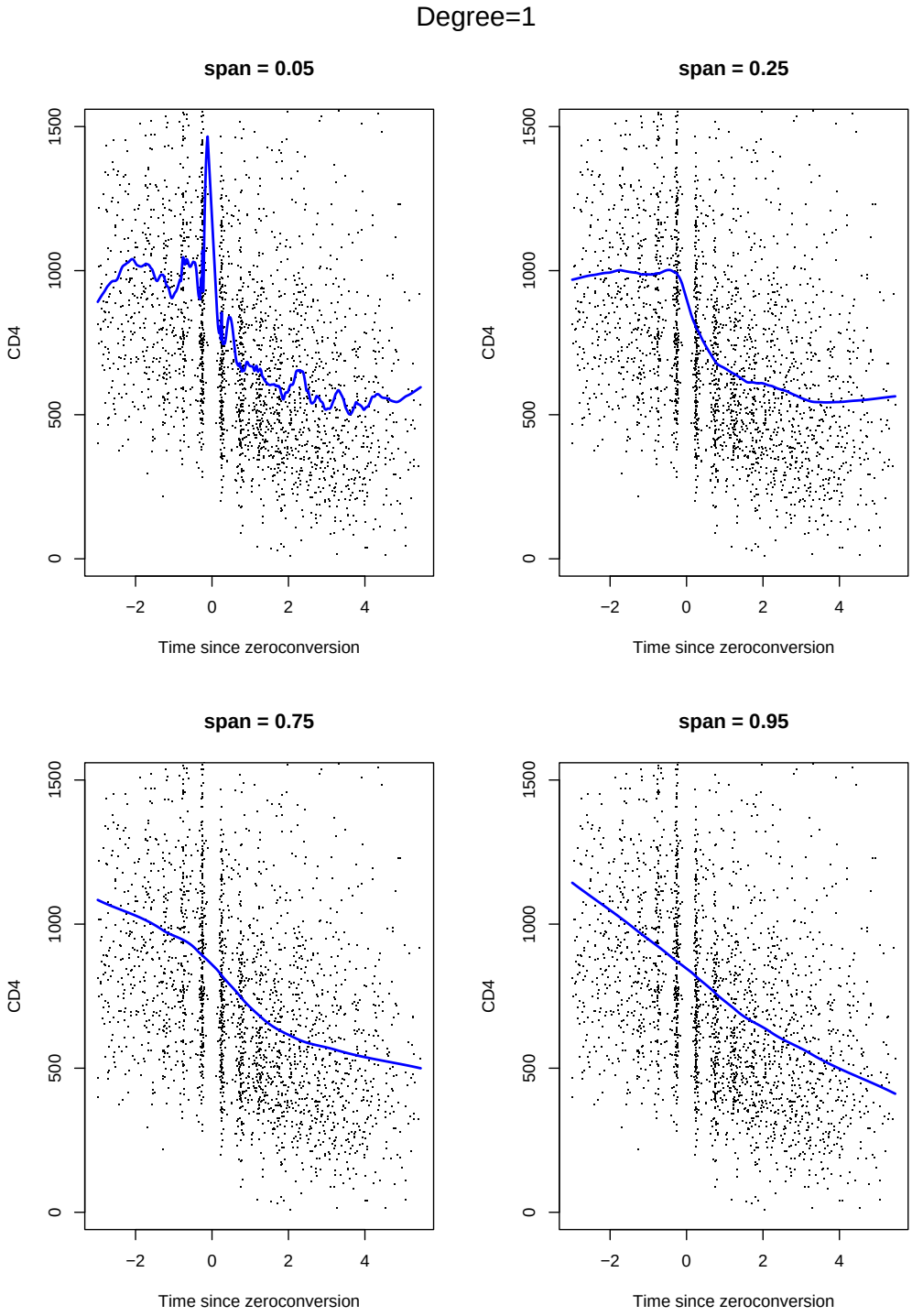


Figure 3.2: CD4 cell count since seroconversion for HIV infected men.

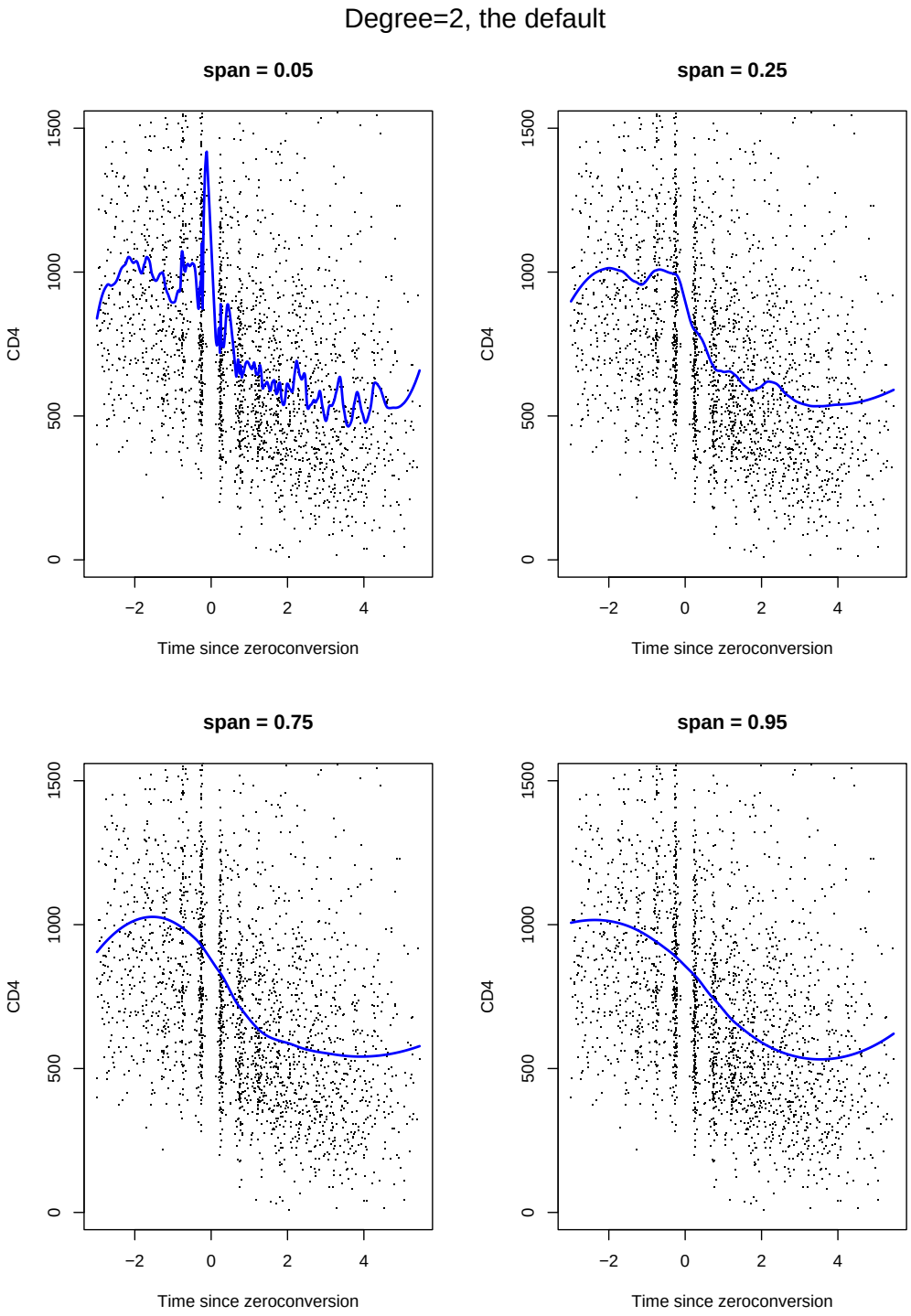
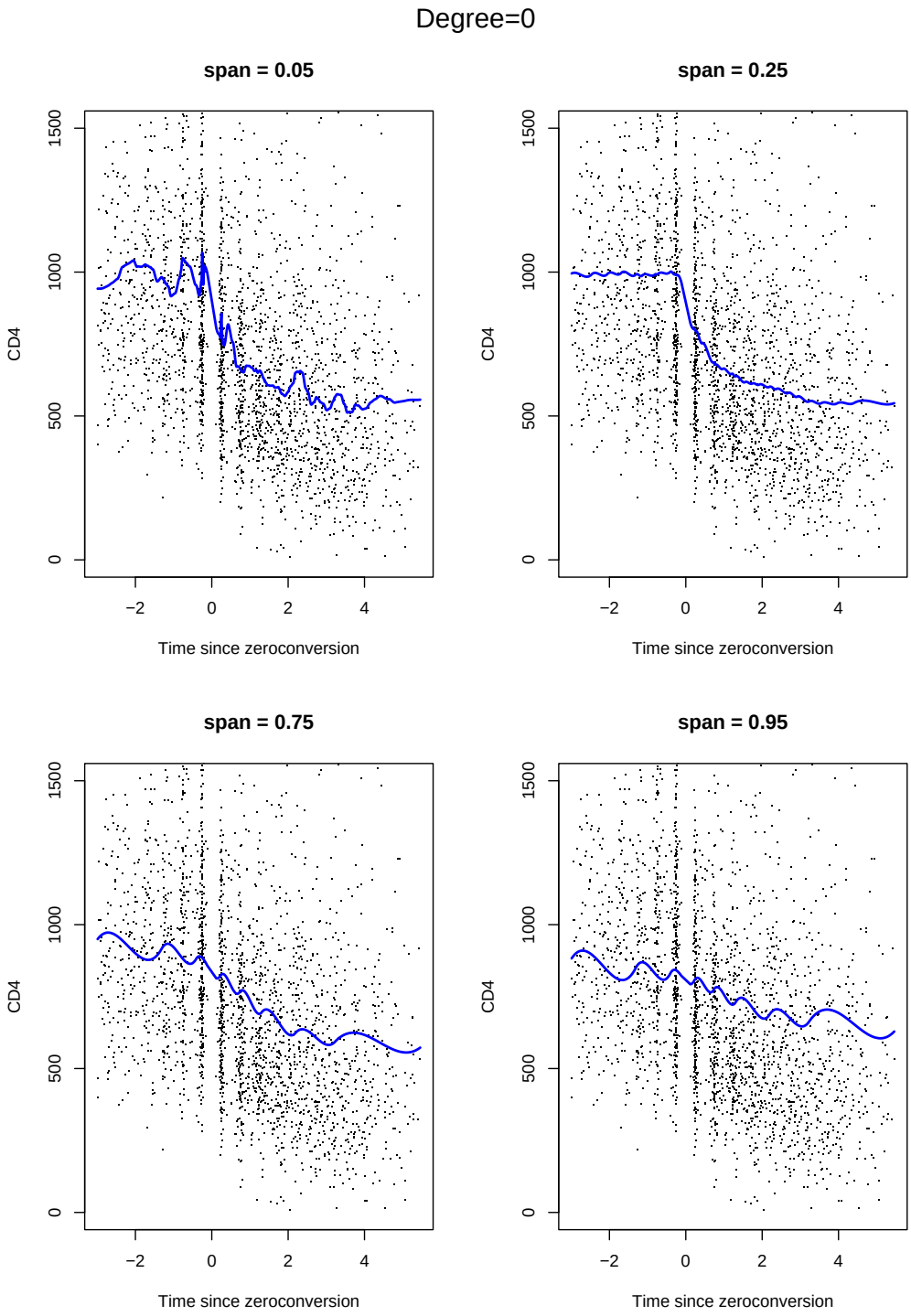


Figure 3.3: CD4 cell count since seroconversion for HIV infected men.



3.4 Symmetric errors and Robust fitting

If the errors have a symmetric distribution (with long tails), or if there appears to be outliers we can use robust loess.

We begin with the estimate described above $\hat{f}(x)$. The residuals

$$\hat{\varepsilon}_i = y_i - \hat{f}(x_i)$$

are computed.

Let

$$B(u;b) = \begin{cases} \{1-(u/t)^2\}^2 & |u| < b \\ 0 & u \geq |u| \geq b \end{cases}$$

be the bisquare weight function. Let m = median($|\hat{\varepsilon}_i|$). The robust weights are

$$r_i = B(\hat{\varepsilon}_i; 6m)$$

The local regression is repeated but with new weights $r_i w_i(x)$. The robust estimate is the result of repeating the procedure several times.

If we believe the variance var$(\varepsilon_i) = a_i\sigma^2$ we could also use this double-weight procedure with $r_i = 1/a_i$.

3.4.1 Example

Radiolabeling based gene expression measurements are useful for cancer research because they can be carried out using small amounts of biological materials. Statistical issues are different from fluorescence expression data, because radiolabeling gives absolute intensities that reflect gene expression and there is no internal control.

The data-set described here was obtained to identify genes that may be associated with lung cancer. Lung cancer tissue was obtained from various subjects. Normal

tissues from the same type of cells was obtained from those same subjects. From each of these tissues 2 samples were prepared using 2 different isotopic batches. Each of these 4 samples were hybridized with a filter spotted with cDNA from many genes in a 48×24 grid. We refer to these spotted filters as arrays. Each of these arrays were scanned to produce an image file which was then analyzed with specialized software that produced an intensity level for each grid point or *spot* on the array.

Not all the values read from the arrays are associated with genes. There were 207 spots where no cDNA was spotted. They were left empty. Because there is *non-specific* binding between the samples and the filters, positive values are obtained from these empty spots. The intensities read from these empty spots provide direct evidence about measurement error associated with the system. Spots associated with genes that are not expressed will also have intensities due to non-specific binding.

Can we rank genes by differential expression between cancer and normal tissues in each subject?

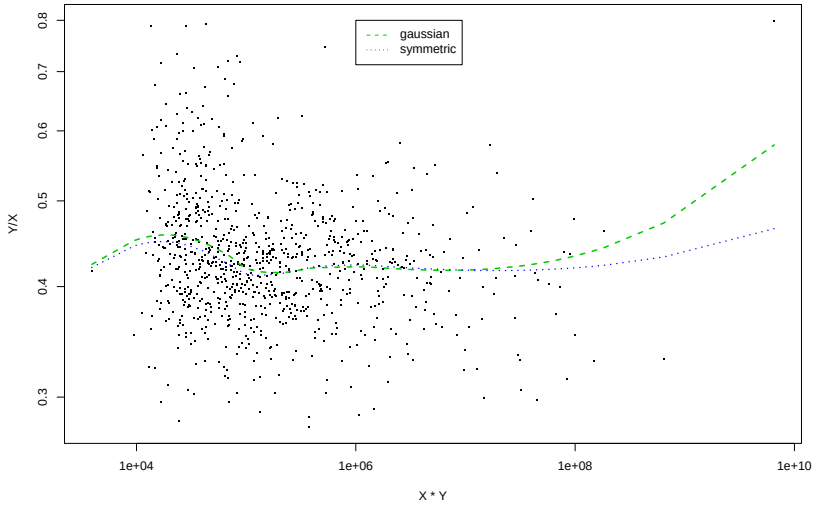
If we denote with $\mathbf{x}$ and $\mathbf{y}$ the log intensities of each spot we could say a gene is differentially expressed if $\mathbf{y} - \mathbf{x}$ is significantly bigger than 0 for the spot related to that gene. One problem with this is that there is a filter effect, so $\mathbf{y}$ can be systematically smaller than $\mathbf{x}$.

A common procedure in microarray data analysis is to simply normalize the filters by subtracting the mean of each filter from each value, i.e. consider $y_i^{(normalized)} = y_i - \bar{y}$ and similarly for the xs. The danger with doing this is that many of the genes spotted on the arrays are usually selected because researchers consider them likely to be over-expressed. This means that the mean of the ys should be larger than the xs and this difference in mean is confounded with the difference in filter effect. By subtracting means we would be subtracting out some of the differential expression between cancer and normal tissues.

In Figure 3.4 we plot the ratio of the intensities vs. the product of the intensities in a log scale, i.e. $y - x$ vs. $x + y$, for the two replicates of subject 1. Notice that the *filter effect* seems to change with the total intensity of a particular spot. For this

reason using medians or trimmed means to remove the filter effect is not a good solution. If we model x and y as random variables then we have that the expected filter effect depends on the total intensity, i.e. $\mathrm{E}(y - x|x + y)$ is not constant. This arises because specific binding and non-specific binding are two different natural processes. Because we have no way of knowing which points represent non-specific binding and which represent specific binding we cannot normalize by just estimating two means. Rather, we estimate $\mathrm{E}(y - x|y + x)$ using loess. It is critical to use a robust loess, so that large differences do not affect the fit too much. Notice in Figure **??** the difference in the robust and non-robust estimates.

Figure 3.4: Total intensity plotted against ratio with a loess prediction using Gaussian and symmetric kernel.



3.5 Multivariate Local Regression

Because Taylor's theorems also applies to multidimensional functions it is relatively straight forward to extend local regression to cases where we have more than one covariate. For example if we have a regression model for two covariates

$$Y_i = f(x_{i1}, x_{i2}) + \varepsilon_i$$

with $f(x, y)$ unknown. Around a target point $\mathbf{x}_0 = (x_{01}, x_{02})$ a local quadratic approximation is now

$$f(x_1, x_2) = \beta_0 + \beta_1(x_1 - x_{01}) + \beta_2(x_2 - x_{02}) + \beta_3(x_1 - x_{01})(x_2 - x_{02}) + \frac{1}{2}\beta_4(x_1 - x_{01})^2 + \frac{1}{2}\beta_5(x_2 - x_{02})^2$$

Once we define a distance, between a point $\mathbf{x}$ and $\mathbf{x}_0$, and a span h we can define define waits as in the previous sections:

$$w_i(\mathbf{x}_0) = W\left(\frac{||\mathbf{x}_i, \mathbf{x}_0||}{h}\right).$$

It makes sense to re-scale x_1 and x_2 so we smooth the same way in both directions. This can be done through the distance function, for example by defining a distance for the space $\mathbb{R}^d$ with

$$||\mathbf{x}||^2 = \sum_{j=1}^{d} (x_j/v_j)^2$$

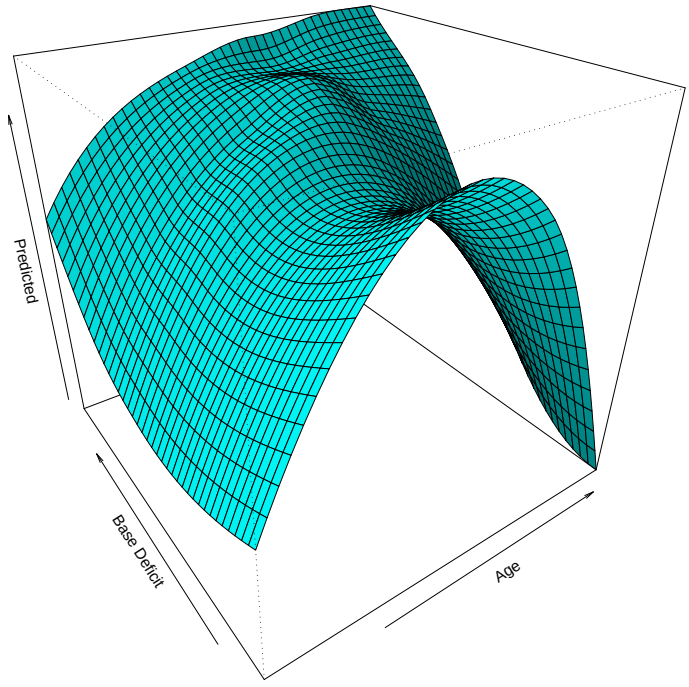
with v_j a scale for dimension j. A natural choice for these v_j are the standard deviation of the covariates.

Notice: We have not talked about k-nearest neighbors. As we will see in Chapter VII the *curse of dimensionality* will make this hard.

3.5.1 Example

We look at part of the data obtained from a study by Socket et. al. (1987) on the factors affecting patterns of insulin-dependent diabetes mellitus in children. The objective was to investigate the dependence of the level of serum C-peptide on various other factors in order to understand the patterns of residual insulin secretion. The response measurement is the logarithm of C-peptide concentration (pmol/ml) at diagnosis, and the predictors are age and base deficit, a measure of acidity. In Figure 3.5 we show a loess two dimensional smooth. Notice that the effect of age is clearly non-linear.

Figure 3.5: Loess fit for predicting C.Peptide from Base.deficit and Age.



Bibliography

[1] Cleveland, R. B., Cleveland, W. S., McRae, J. E., and Terpenning, I. (1990). Stl: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, 6:3–33.

[2] Cleveland, W. S. and Devlin, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83:596–610.

[3] Cleveland, W. S., Grosse, E., and Shyu, W. M. (1993). Local regression models. In Chambers, J. M. and Hastie, T. J., editors, *Statistical Models in S*, chapter 8, pages 309–376. Chapman & Hall, New York.

[4] Loader, C. R. (1999), *Local Regression and Likelihood*, New York: Springer.

[5] Socket, E.B., Daneman, D. Clarson, C., and Ehrich, R.M. (1987). Factors affecting and patterns of residual insulin secretion during the first year of type I (insulin dependent) diabetes mellitus in children. *Diabetes* 30, 453–459.